Ramez Abdel Monem Shendy Student Code: **201720045**

NNets2020

Course: Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization

Week 2,3 Summary page

This course in this deep learning specialization concerned about how to improve the neural network performance, by applying different types of machine learning and deep learning techniques to improve the dynamics of the deep learning model.

This week's (week 2 and 3) main concepts are:

1- Batch Vs. mini-batch gradient descent
2- Exponentially weighted averages
3- Momentum
4- RMSProp
5- Adam
6- Learning rate decay
7- optimization problems
8- Hyperparameter tuning
9- Batch Normalization

1- Batch Vs. mini-batch gradient descent:

When the input dataset is very large and sometimes millions or records and in images its hundreds of features, it'll be very slow to train the model using batch gradient descent as for the gradient descent algorithm to take one step towards convergence it'll use the whole dataset in each iteration. That is why mini-batch gradient descent is used and sometimes stochastic gradient descent, the mini batch is considered as a hyperparameter to tune in the network and 64, 128, 256, … input dimensions can be used or search in those to find the best input dimension.
Stochastic gradient descent will only use one training example in the training in each iteration, mini-batch will use n training examples and batch will use the m examples.

2- Exponentially weighted averages:

Statistical concept to smooth line drawn with many points with high oscillations, Faster than basic gradient descent when introduced in the optimization algorithm, also known as exponentially weighted moving averages in statistics.
Has risen to solve the problem of oscillating cost over iterations in mini-batch gradient descent

3- Momentum:

Exponentially weighted average of the gradients.
Reduces oscillations without the need of increasing the learning rate.

4- RMSProp:

Squaring the derivatives and taking the square root.
Allows increasing the learning rate without the risk of divergence.
Introduced by Geofrrey Hinton in a neural network course at coursera.

5- Adam:

Momentum and RMSprop combined in one algorithm.
Adaptive moment estimation, widely used and accepted, also gives good results and robust across various deep learning applications.

6- Learning rate decay:

Can be done manually by watching and governing the model while learning and decide to decrease the learning rate alpha as learning slows down or by using the learning rate decay equation to automate the process a little bit.

7- Optimization problems:

local optima and saddle points in optimization.
The probability to face saddle points are much higher to face than local minima in higher dimensions when training a deep neural network.
Platueaus Can really slow down learning, it’s the region where the derivative is very close to zero for a long time.
These problems can be pretty much solved by algorithms like RMSProb, Momentum and Adam.

8- Hyperparameter tuning:

Adjusting neural network various hyperparameters can be the difference between a good model and a bad one specially if we considered the number of neurons or the number of layers in the model architecture.
The most important hyperparameters in a neural network model are: the learning rate alpha, if using adam or momentum we then have to cope with theit hyperparameters as well and maybe the mini batch size.
Tuning hyperparameters can be handled by different approaches, one can be babysitting the model overtime and see how it behaves. The other one is training models in parallel and see which is better at the end.

9- Batch normalization:

Similar to input normalization but done at the layers after the input layer normalizing the output of neurons which makes the later weights in the network more robust to change s to weights in earlier layers.